第28卷 第1期 工 程 数 学 学 报 Vol. 28 No. 1
2011年02月 CHINESE JOURNAL OF ENGINEERING MATHEMATICS Feb. 2011

文章编号:1005-3085(2011)01-0001-06

基于支持向量机的青年女性红细胞压积参考值与地理环境关系研究*

刘 倩[1]， 刘三阳[1]， 葛 淼[2]

(1- 西安电子科技大学数学系，西安 710071； 2- 陕西师范大学地理系，西安 710062)

摘 要: 为制定中国青年女性红细胞压积参考值的统一标准提供科学依据，收集中国各地用温氏法测定的健康青年女性红细胞压积参考值，提出了一种基于支持向量机的分析青年女性红细胞压积参考值与地理环境的海拔高度、年日照时数、年平均相对湿度、年平均气温、年降水量等要素关系的方法．知道了中国某地的地理因素，就可以用此模型预测这个地区的青年女性红细胞压积参考值．实验结果表明，支持向量机预测模型能够克服地理要素本身之间具有的共线性，反映红细胞压积参考值与地理环境之间的非线性、复杂性关系特征，并且保持一定的预测精度，与主成分回归模型相比，更适应于红细胞压积参考值的预测．

关键词: 支持向量机；红细胞压积；参考值；地理环境；血液流变学

分类号: **AMS**(2000) 68Q32 **中图分类号**: R188; O212.4 **文献标识码**: A

1 引言

红细胞压积是血液流变学的一个重要指标．目前，国内外缺乏血液流变学指标参考值的统一标准，影响了临床和科研成果的准确性和可比性．为制定青年女性红细胞压积参考值的统一标准提供科学依据，很多人测定了本地区的青年女性红细胞压积参考值[1-3]．但对不同地区红细胞压积参考值之间的关系以及红细胞压积参考值与地理环境之间的关系研究较少．

红细胞压积参考值与正常人生活的地理环境关系非常密切，但地理环境对红细胞压积参考值的影响是通过人们所处的大气条件、饮食结构、居住环境等要素作用于人体的血液而进行的．对地理环境与红细胞压积参考值的这种复杂的非线性关系的研究具有非常重要的意义．文献 [4-6] 用相关分析、回归分析的方法研究了红细胞压积参考值与地理环境的关系，文献 [7] 用主成分回归的方法研究了中年男性血红蛋白参考值与地理因素的关系．由于回归分析和主成分回归反映的仅是变量之间的线性关系，所以比较难以体现血液流变学的重要指标与地理环境之间复杂的非线性关系．同时回归分析方法本身要求自变量间线性无关，由于地理要素本身之间具有较大的相关性，这也影响了模型的预测精度．

针对这种情况，文献 [8] 应用人工神经网络技术进行非线性建模．其基本原理是通过大样本的自我学习来映射输入和输出关系，从而达到预测控制的目的．然而此模型还存在某些难以解决的问题，例如需要大量的数据样本以及模型存在过拟合现象等，这在很大程度上制约着该技术在红细胞压积参考值与地理环境关系的研究中的应用．

收稿日期: 2008-03-10. **作者简介**: 刘倩 (1979年8月生)，女，讲师．研究方向：地理数据挖掘与生物信息学．
***基金项目**: 国家自然科学基金 (49771007).

近几年，一种基于由Vapnik提出的统计学习理论[9] 发展而来的支持向量机(support vector machine, 即SVM)算法正受到越来越多的关注，其卓越性能体现在：与人工神经网络类似，SVM也是一个完全基于数据的非线性建模工具；SVM模型基于结构风险最小化原则，泛化性能潜力巨大；SVM的目标函数是一个凸优化问题，其最优解具有唯一性；在SVM模型中，应用核技术，将输入空间中的非线性问题通过非线性函数映射到高维特征空间中，在高维空间中构造线性判别函数；SVM专门针对小样本情况，其最优解基于已有样本信息，而不是样本数趋于无穷大时的最优解．现有的文献中较少发现有SVM模型用于研究红细胞压积参考值与地理环境的关系．本文拟就此展开讨论．

本文首先介绍资料来源，进而对中国各地用温氏法(wintrobe)测定的139组健康青年女性红细胞压积参考值与地理环境的数据进行相关分析，其次介绍SVM算法，提出了基于支持向量回归的研究青年女性红细胞压积参考值与地理环境之间关系的方法，发现有一定的规律性．

2 资料

收集了中国139个市(县)级医院和有关研究单位及高等院校测定的8486例健康青年女性红细胞压积参考值；年龄范围是18-25岁之间的青年女性；这些单位分布在中国31个省、市、自治区，缺乏台湾省、香港特别行政区、澳门特别行政区的资料，东部平原地区的资料多于西部高原地区的资料．测定红细胞压积参考值的方法有多种，本文收集的是用温氏法测定的红细胞压积参考值．温氏法测定的方法是：常规采静脉血2.5ml，注入肝素抗凝试管中，轻轻混匀，用毛细吸管吸取抗凝血慢慢加入温氏压积管内至“10”刻度处，不能有气泡，将压积管放入离心机中，以2300g的离心力离心30min，直到红细胞体积不再改变为止，读取右侧红细胞层的高度，读数乘10即为红细胞压积百分率．

地理环境主要选取的指标是海拔高度x_1，年日照时数x_2，年平均相对湿度x_3，年平均气温x_4，年降水量x_5等五项指标．地理资料中的海拔高度来源于测绘局数据中心提供的共享资料；年平均气温和年降水量、年相对湿度来源于国家气象局数据中心提供的共享资料；年日照时数和共享资料中未列出的数据主要取材于有关地理著作和辞典．

3 相关分析

运用相关分析计算出青年女性红细胞压积参考值y与海拔高度x_1，年日照时数x_2，年平均相对湿度x_3，年平均气温x_4，年降水量x_5的简单相关系数r分别是

$$r_{x_1,y}=0.908,\quad r_{x_2,y}=0.507,\quad r_{x_3,y}=-0.678,\quad r_{x_4,y}=-0.863,\quad r_{x_5,y}=-0.594.$$

分析结果显示，红细胞压积参考值与各个地理因素的相关性在0.01的显著性水平下显著，同时，各个地理因素之间也具有较强的共线性．

4 基于支持向量机的回归算法

假设训练样本集

$$G=\{\boldsymbol{x}_i,y_i\}_{i=1}^{N},$$

其中 $\boldsymbol{x}_i \in \mathbf{R}^m$ 为输入值，$y_i \in \mathbf{R}$ 为输出值，支持向量回归模型的基本思想就是将 m 维输入向量 $\boldsymbol{x}$ 通过某种非线性关系 ϕ 映射到高维特征空间 F 中，从而在特征空间 F 中实现线性回归

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i \phi^T(\boldsymbol{x}_i)\phi(\boldsymbol{x}) + b, \tag{1}$$

其中 $\phi(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 在特征空间上的映射；未知参数 $\{\omega_i\}_{i=1}^N$ 和 b 分别表示权重和偏差系数，可以在样本集训练过程中获得，为了避免出现过拟合现象进而提高模型的泛化能力，需要考虑结构风险原则并使下列函数极小化

$$R(\omega) = \frac{1}{N}\sum_{i=1}^{N} \left|f(\boldsymbol{x}_i) - y_i\right|_\varepsilon + \lambda\|\boldsymbol{\omega}\|^2, \tag{2}$$

其中 λ 为调解因子，$|f(\boldsymbol{x}_i) - y_i|_\varepsilon$ 定义为 Vapnik-ε 不敏感损失函数，其表达式为

$$\left|f(\boldsymbol{x}_i) - y_i\right|_\varepsilon = \begin{cases} |f(\boldsymbol{x}_i) - y_i| - \varepsilon, & |f(\boldsymbol{x}_i) - y_i| \geq \varepsilon, \\ 0, & |f(\boldsymbol{x}_i) - y_i| < \varepsilon, \end{cases} \tag{3}$$

其中 $f(\boldsymbol{x})$ 为通过对样本集的学习而构造的回归估计函数，y 为对应的 $\boldsymbol{x}$ 目标值，$\varepsilon > 0$ 为与函数估计精度直接相关的设计参数，将 ε 不敏感损失函数形象地比喻为 ε 管道，它意味着不惩罚偏差小于 ε 的误差项.

Vapnik[9] 认为式 (2) 极小化后可以得到

$$f(\boldsymbol{x}, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b, \tag{4}$$

其中 α_i, $\alpha_i^* \geq 0$ 为拉格朗日乘子，$K(\boldsymbol{x}_i, \boldsymbol{x})$ 为核函数且有

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi^T(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j). \tag{5}$$

式 (4) 和 (5) 的一个重要特点是对于特征 $\phi(\boldsymbol{x})$，核函数 K 都可以被解析表达且形式相对简单，因此，无需将矢量 $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ 直接映射到特征空间 F 中，即无需计算 $\phi(\boldsymbol{x}_i)$, $\phi(\boldsymbol{x}_j)$，就可以计算特征空间的内积，但是前提是该核函数必须满足 Mercer 条件. 常见的核函数有多项式、径向基以及 Sigmoidal 函数等. 根据凸优化的充要条件，拉格朗日乘子 α_i, α_i^* 可由下式获得

$$\begin{aligned} &\max\ \ R(\alpha_i, \alpha_i^*) = -\frac{1}{2}\sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* + \alpha_i) \\ &\text{s.t.}\ \ \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i, \alpha_i^* \leq C. \end{aligned} \tag{6}$$

由上述优化方程，可以求出 α_i, α_i^*. 对于 Vapnik-ε 不敏感损失函数而言，拉格朗日乘子 α_i, α_i^* 具有稀疏性，即只有 α_i, α_i^* 不为零所对应的向量 $\boldsymbol{x}$ 被称为支持向量、称式 (4) 为支持向量回归模型. 从上面的论述中，不难发现与常规的回归方法相比，支持向量回归模型具有两方面的优势：采用结构风险最小化作为优化目标，提高了回归函数的泛化能力；引入了核方法，实现了低维数据空间与高维特征空间的非线性映射，提高了回归函数的非线性数据处理能力. SVM 回归算法详细的描述可以参考文献 [10]. 基于此，本文用 SVM 回归模型来分析和预测青年女性红细胞压积参考值与地理环境复杂的非线性关系.

5 青年女性红细胞压积参考值预测模型

5.1 模型设计

在青年女性红细胞压积参考值与地理要素的关系研究中，主要构建青年女性红细胞压积参考值的预测模型．在预测型学习任务中，模型及其参数选择是有效预测的前提，这里采用ε-支持向量回归机作为青年女性红细胞压积参考值的预测模型．该模型中需要确定的结构参数有：

1) 惩罚因子C：表示在决策函数的复杂性和决策误差之间的折中程度；

2) ε：表示ε不敏感损失函数中的偏差；

3) 核函数：这里采用较为常用的高斯径向基函数

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\big(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2\big),$$

其中σ是核函数的带宽．

5.2 参数选择

模型选定后，需要人为确定的参数包括(ε,C,σ)．这些参数一般无法直接获得，而且这些参数的确定没有严格的理论作指导，一定程度上依赖于使用者的经验和试凑与比较．这里假定$\varepsilon=0.1$, $C=1.7$, $\sigma=0.37$.

5.3 试验和分析

利用支持向量机方法进行青年女性红细胞压积参考值的预测，首先要确定影响青年女性红细胞压积参考值的主要地理要素，其次要选择具有观测资料的地理要素和青年女性红细胞压积参考值构成样本数据集，然后利用SVM进行学习训练，最后根据训练所得参数进行预测．

选择与青年女性红细胞压积参考值y相关性显著的地理要素：海拔高度x_1，年日照时数x_2，年平均相对湿度x_3，年平均气温x_4，年降水量x_5作为ε-SVR的输入向量，青年女性红细胞压积参考值y作为该模型的输出变量．采集中国各地用温氏法测定的139组健康青年女性红细胞压积参考值以及相应的地理要素的数据作为样本数据集，这139组数据是对收集的中国139个市(县)级医院和有关研究单位及高等院校测定的8486例健康青年女性红细胞压积参考值进行均值预处理之后得到的．其中，前133组用于模型训练，后6组数据用于模型测试，这六组测试数据分别取自中国的青藏区，西南区，西北区，东南区，华北区，东北区，代表的城市分别为拉萨、贵阳、银川、南昌、北京、长春．

目前，基于SVM的算法软件已经相对成熟，本文模型训练采用LIBSVM 2.8软件包，该软件包主要应用SMO算法求解凸优化问题，具有快速高效的特点．基本步骤如下：

1) 为了消除由于量纲和单位不同造成的影响，并且避免在训练时计算核函数时引起数值计算的困难，对样本的输入、输出数据用svmscale.exe程序分别进行归一化处理，数据略；

2) 确定模型的输入、输出关系后，假定参数$\varepsilon=0.1$, $C=1.7$, $\sigma=0.37$，用svmtrain.exe对133组样本数据进行训练，获得模型；

3) 用svmpredict.exe对6组测试数据进行测试，表1给出了训练完成后模型测试试验的结果，通过相关系数R，差值均平方MSE和平均相对误差$e\%$等统计指标与对应的主成分回归模型进行了对比．

可以看到，就主成分回归模型而言，支持向量回归模型的预测精度与可靠性均有明显提高，由于受数据资料的影响，青年女性红细胞压积参考值是一个比较难以预测的指标，但即使这样，支持向量回归模型的预测能力同主成分回归模型相比较也得到了一定程度的改善．

表 1: 青年女性红细胞压积参考值预测模型性能对比

样本点	x_1	x_2	x_3	x_4	x_5	实测值 y	主成分回归预测值 $\hat{y}$	支持向量机回归预测值 $\hat{y}$
拉萨	3658.0	3007.7	45	7.5	454.0	50.5	48.8	48.7
贵阳	1071.2	1371.0	79	15.3	1174.7	42.1	42.8	42.3
银川	1111.5	3039.6	59	8.5	202.8	43	43.7	44.1
南昌	46.7	1903.9	77	17.5	1596.4	40.5	38.9	40.2
北京	31.2	2780.2	60	11.5	644.2	41	40.7	41.3
长春	236.8	2643.5	65	4.9	593.8	41.3	43.7	41.9
R							0.908	0.978
MSE							2.05	0.85
e%							3	1.2

6 结论

从简单相关系数可以看出，青年女性红细胞压积参考值与地理环境的海拔高度、年日照时数、年平均相对湿度、年平均气温和年降水量的相关性很显著，其中海拔高度是影响青年女性红细胞压积参考值最主要的因素．研究表明：随着海拔高度的逐渐增大，空气逐渐稀薄，氧含量逐渐减小，机体为了适应缺氧的环境，血液中的红细胞数代偿性的逐渐增加，导致青年女性红细胞压积参考值的逐渐增大．

如果知道了中国某地的海拔高度、年日照时数、年平均相对湿度、年平均气温和年降水量等地理因素指标，就可以用建立的支持向量回归预测模型来估算这个地区的青年女性红细胞压积参考值．例如：北京的海拔高度是31.2m，年日照时数是2780.2h，年平均相对湿度是60.0%，年平均气温是11.5℃，年降水量是644.2mm，用支持向量回归模型计算得青年女性红细胞压积参考值的预测值为41.3．

基于主成分回归的红细胞压积参考值预测模型是先对地理因素进行线性组合，得到彼此之间不相关的新的综合变量，再将综合变量与红细胞压积参考值进行回归分析，可以剔除重叠的信息从而使问题得到综合简化．而基于支持向量机方法的预测模型由于采用结构风险最小化作为优化目标，从而提高了回归函数的非线性数据处理能力，在“噪音”数据环境下仍能保持一定的预测精度．实验数据表明，该方法克服了地理要素本身之间具有的共线性，反映了红细胞压积参考值与地理环境之间的非线性、复杂性关系特征，与主成分回归模型相比，更适应于红细胞压积参考值的预测．

致谢：裴树萱、孙志新、刘崇礼、李卫兵、宋玉舒等同志提供了数据资料，特此致谢．

参考文献：

[1] 杜智敏，刘崇礼，郑永梅．高原地区不同移居时间正常成人血液流变学测定及分析[J]．微循环技术杂志，1994，2(3)：134-135
Du Z M, Liu C L, Zheng Y M. The measurement and analysis of hemorheology of different time to normal adult in plateau area[J]. Journal of Chinese Microcirculation, 1994, 2(3): 134-135

[2] 路遥，刘娟英. 哈尔滨市血液流变学指标正常参考值调查分析[J]. 中国血液流变学杂志，1997, 7(3): 21-23
Lu Y, Liu J Y. The investigation and analysis of normal reference value of hemorheology in Harbin[J]. Chinese Journal of Hemorheology, 1997, 7(3): 21-23
[3] 韦丽华，高宗鹰，马爱国. 昆明市血液流变学指标正常参考值调查[J]. 中国血液流变学杂志，1999, 9(1): 49-51
Wei L H, Gao Z Y, Ma A G. The investigation of normal reference value of hemorheology in Kunming[J]. Chinese Journal of Hemorheology, 1999, 9(1): 49-51
[4] 葛淼. 中国中老年人红细胞压积参考值与地理因素的关系[J]. 地理科学，1999, 19(1): 78-81
Ge M. Relationship between reference values of Chinese adults and old peoples hematocrit and geographical factors[J]. Scientia Geographica Sinica, 1999, 19(1): 78-81
[5] 葛淼. 中国中老年人红细胞压积与地理因素的逐步回归分析[J]. 航天医学与医学工程，1999, 12(1): 37-41
Ge M. Stepwise regression analysis of hematocrit value in Chinese adult and old people with geographical factors[J]. Space Medicine and Medical Engineering, 1999, 12(1): 37-41
[6] Ge M. The relationship between reference value of erythrocyte sedimentation rate and geographical factors[J]. Bioscience Report, 2001, 21(3): 287-292
[7] 刘倩，刘三阳，葛淼. 中年男性血红蛋白参考值与中国地理因素分析[J]. 西北大学学报，2008, 38(3): 494-497
Liu Q, Liu S Y, Ge M. Analysis between reference value of middle aged men's hemoglobin and geographical factors in China[J]. Journal of Northwest University, 2008, 38(3): 494-497
[8] 杨青生，张红贤，葛淼. 基于人工神经网络的老年男性血沉参考值与地理环境关系研究[J]. 地理科学，2006, 26(6): 749-754
Yang Q S, Zhang H X, Ge M. Relationship between reference value of Chinese old men's erythrocyte sedimentation rate (ESR) and geographical factors based on NN[J]. Scientia Geographica Sinica, 2006, 26(6): 749-754
[9] Vapnik V N. Statistical Learning Theory[M]. New York: Wiley & Sons, 1998
[10] 郭崇慧，陆玉昌. 预测型数据挖掘中的优化方法[J]. 工程数学学报，2005, 22(1): 25-29
Guo C H, Lu Y C. Optimization methods in predictive data mining[J]. Chinese Journal of Engineering Mathematics, 2005, 22(1): 25-29

Research on Relationship Between Reference Value of Young Women's Hematocrit and Chinese Geographical Factors Based on Support Vector Machines

LIU Qian[1], LIU San-yang[1], GE Miao[2]

(1- Department of Mathematics, Xidian University, Xi'an 710071;
2- Department of Geography, Shaanxi Normal University, Xi'an 710062)

Abstract: In order to provide a basis for unifying the reference value criteria standard of Chinese young women's hematocrit, the paper discusses the nonlinear relationship between the reference value of Chinese healthy young women's hematocrit which are determined by the wintrobe laws and geography factors based on support vector machines. The selected geographical factors include altitude, annual sunshine hour, annual average relative humidity, annual average temperature and annual precipitation. If the geographical values are obtained in some area, the reference value of Chinese young women's hematocrit of this area can be reckoned using this SVM model. Experimental results show that the SVM model is capable of overcoming the multicollinearity between the geography factors and maintaining the stability of the predictive accuracy, and is more suitable for predicting the hematocrit value than the principal component regression analysis.

Keywords: support vector machines; hematocrit; reference value; geographical environment; hemorheology

Received: 10 Mar 2008. **Accepted:** 19 Oct 2010.
Foundation item: The National Natural Science Foundation of China (49771007).